Научно-исследовательский журнал «Modern Economy Success»
https://mes-journal.ru
2025, № 5 / 2025, Iss. 5 https://mes-journal.ru/archives/category/publications
Научная статья / Original article
Шифр научной специальности: 5.2.2. Математические, статистические и инструментальные методы в экономике (экономические науки)
УДК 330.46

[1] Панкова Л.В., [1] Старченкова О.Д.,
[1] Санкт-Петербургский политехнический университет Петра Великого

Применение алгоритма акторного доминирования генов при определении параметров анализа информационной среды экономической деятельности предприятия

Аннотация: актуальность данного исследования обусловлена необходимостью разработки и внедрения интеллектуальных алгоритмов анализа информационной среды экономической деятельности предприятия. В условиях цифровой экономики организации функционируют в сложной и динамичной информационной среде, оказывающей прямое влияние на их эффективность и конкурентоспособность. Предлагаемый в статье алгоритм акторного доминирования генов направлен на повышение точности и эффективности анализа ключевых параметров этой среды. Цель исследования: сформировать инструмент анализа информационной среды экономической деятельности предприятия (в дальнейшем – инструмент взаимодействия с информационной средой) в разрезе выявления сущностно-содержательных компонент текстовой информации, с помощью методов квантификации и построения параметрической структуры ограничений алгоритмики в рамках управления качеством работы методологии. Применение данного подхода позволяет выявлять доминирующие информационные потоки, определять влияние внешних и внутренних акторов на экономическую деятельность, а также формировать обоснованные управленческие решения на основе обработки больших данных. Результаты исследования способствуют совершенствованию методов анализа информационного взаимодействия в цифровом пространстве и автоматизации процессов мониторинга и оценки состояния информационной среды предприятия. Работа выполнена в рамках реализации проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008).

Ключевые слова: информационная среда предприятия, акторное доминирование генов, генетический алгоритм, кластеризация текстовой информации, сущностно-содержательные компоненты, параметрическое сочетание, Latent Dirichlet Allocation (LDA), коэффициент силуэта, цифровая экономика, анализ предпочтений аудитории

Для цитирования: Панкова Л.В., Старченкова О.Д. Применение агоритма акторного доминирования генов при определении параметров анализа информационной среды экономической деятельности предприятия // Modern Economy Success. 2025. № 5. С. 207 – 219.

Поступила в редакцию: 10 июня 2025 г.; Одобрена после рецензирования: 8 августа 2025 г.; Принята к публикации: 23 сентября 2025 г.

[1] Pankova L.V., [1] Starchenkova O.D.,
[1] Peter the Great St. Petersburg Polytechnic University

The application of the actor dominance gene algorithm in determining the parameters for information environment analysis

Abstract: the relevance of this study is driven by the need to develop and implement intelligent algorithms for analyzing the information environment of a company's economic activity. In the context of the digital economy, organizations operate within a complex and dynamic information environment that directly affects their efficiency and competitiveness. The actor-based gene domination algorithm proposed in this article is aimed at improving the accuracy and effectiveness of analyzing the key parameters of this environment. The purpose of the research is to develop a tool for analyzing the information environment of a company's economic activity (hereinafter referred to as the information environment interaction tool) by identifying the essential semantic components of textual information using methods of quantification and constructing a parametric structure of algorithmic constraints within the framework of quality management methodology. The application of this approach enables the identification of dominant information flows, the assessment of the influence of internal and external actors on economic activity, and the formation of well-founded managerial decisions based on big data processing. The results of the study contribute to the advancement of methods for analyzing information interactions in the digital space and to the automation of monitoring and evaluation processes regarding the state of a company's information environment. The work was carried out within the framework of the project "Development of a methodology for the formation of an instrumental base for the analysis and modeling of spatial socio-economic development of systems in the context of digitalization based on internal reserves" (FSEG-2023-0008).

Keywords: information environment of the enterprise, actor dominance of genes, genetic algorithm, text information clustering, content-based components, parametric combination, Latent Dirichlet Allocation (LDA), silhouette coefficient, digital economy, audience preference analysis

For citation: Pankova L.V., Starchenkova O.D. Algorithmics for aggregating unique digital identifiers of potential applicants localized within a regional agglomeration. Modern Economy Success. 2025. 5. P. 207 – 219.

The article was submitted: June 10, 2025; Approved after reviewing: August 8, 2025; Accepted for publication: September 23, 2025.

Введение

Цель исследования: сформировать инструмент анализа информационной среды экономической деятельности предприятия (в дальнейшем – инструмент взаимодействия с информационной средой) в разрезе выявления сущностно-содержательных компонент текстовой информации, с помощью методов квантификации и построения параметрической структуры ограничений алгоритмики в рамках управления качеством работы методологии.

Ключевой задачей является формирование этапов алгоритма акторного доминирования генов и определение параметрической структуры ограничений для анализа информационной среды экономической деятельности предприятия с целью выявления параметрического сочетания, обеспечивающего соответствие результатов заданным требованиям к анализу.

Научная новизна исследования заключается в применении акторного доминирования генов для определения параметрического сочетания, формирующего значимые в рамках исследования результаты.

В результате был разработан новый подход к определению параметров анализа информационной среды экономической деятельности предприятия, основанный на применении алгоритма акторного доминирования генов. Он позволил повысить качество обработки данных

Для достижения поставленных целей, авторы использовали методы машинного обучения, которые применялись на нескольких выборках данных, состоящих из открытых профилей интернет-пользователей,отфильтрованных по целевым признакам. Инструмент реализован на языке Python.

В результате исследования был разработан инструмент анализа информационной среды экономической деятельности предприятия, основанный на алгоритме акторного доминирования генов, обеспечивающем автоматизацию обработки больших объемов текстовой информации. Методология включает квантификацию текстов, построение

параметрической структуры ограничений и кластеризацию сущностно-содержательных компонент с использованием LDA и оценки качества через коэффициент силуэта. Генетический алгоритм оптимизирует параметры кластеризации, формируя поколенческую модель с трендом повышения эффективности, что позволяет достигать плато качества анализа. Инструмент прошёл апробацию на примере цифрового взаимодействия вуза с потенциальными абитуриентами, выявляя устойчивые паттерны интересов и сегментируя целевую аудиторию. Это открывает возможности для построения персонализированных коммуникационных стратегий, разработки продуктов, таргетированного маркетинга и профориентации. Полученные результаты могут использоваться для повышения конкурентоспособности и эффективности предприятий за счёт точного анализа цифровой информационной среды. Разработанный подход масштабируем, универсален и применим как в коммерческом, так и в некоммерческом секторах экономики.

Прикладной характер исследования: полученные результаты могут быть полезны для коммерческих и некоммерческих организаций в рамках взаимодействия с цифровой информационной средой с целью повышения эффективности экономической деятельности (в частности, инструмент опробован на примере автоматизации цифрового взаимодействия университета с потенциальными абитуриентами – автоматизация цифрового взаимодействия подразумевает использование программного обеспечения и алгоритмов для механизации различных рутинных задач – в данном примере обеспечивая упрощение этапов первичной коммуникации университета с абитуриентами (например, а рамках приёмной кампании)). Взаимодействие между организациями и цифровой информационной средой представляет собой двусторонний процесс обмена информацией и ресурсами. Организации, в свою очередь, генерируют, распространяют и потребляют информацию в цифровой среде, взаимодействуя с аудиторией, партнерами, клиентами. Цели такого взаимодействия могут быть различными и характеризуются широкой вариативностью: увеличение узнаваемости (распространение информации о компании, ее продуктах и услугах в цифровой среде, формирование положительного имиджа); привлечение аудитории (использование цифровых каналов для привлечения потенциальных клиентов и покупателей); улучшение коммуникации (повышение прозрачности и доступности информации для аудитории, выявление и развитие приоритетных каналов и способов коммуникации); анализ данных (сбор и анализ данных для принятия более эффективных решений, понимания потребностей аудитории). В совокупности организация таких взаимодействий на основе использования алгоритмов способствует повышению эффективности деятельности предприятий и организаций.

Материалы и методы исследований

Размерность потоков цифровой информации, сопровождающей экономическую деятельность предприятий и организаций, требует автоматизации математических алгоритмов обработки и расширения функционала аналитического инструментария [1].

Одним из значимых направлений подходов к анализу, связанных с цифровым характером информации, является взаимодействие с электронными профилями пользователей социальных сетей и организаций. Генерируемый ими текст может предоставлять широкий спектр значимых информационных компонент [2].

В связи с этим, организации, как коммерческие, так и некоммерческие, обладающие инструментами обработки цифровой информации, получают значимое конкурентное преимущество. Оно заключается в возможности более быстрого и качественного анализа данных, что позволяет отимизировать бизнес-процессы, повысить эффективность работы, принимать взвешенные решения и быстрее реагировать на изменения рынка. Доступ к обширным массивам данных позволяет создавать новые продукты и услуги, выходить на новые рынки и более продуктивно строить коммуникацию с клиентами. Поэтому актуальным является формирование соответствующих алгоритмов и инструментов [9].

В данном исследовании предлагается построить инструментарий, позволяющий выявить группы интересов определённой целевой аудитории на примере потенциальных абитуриентов университета. То есть, исследование осуществляется в контексте аналитики предпочтений исследуемой группы. Кроме того, с точки зрения управления выявлять первичные паттерны предпочтений, которые потенциально при соответствующем информационном фоне могут быть развиты в предпочтения к группе наиболее актуальных (в рамках кадровой безопасности / в разрезе популярности направлений обучения) специальностей и направлений подготовки.

Сам по себе анализ предпочтений аудитории, осуществляемый в цифровой среде, позволяет организациям, как коммерческим, так и некоммерческим, определять потребности и интересы аудитории, что является ключевым фактором достижения успеха в конкурентной борьбе. Полученные знания о предпочтениях играют решающую роль в принятии стратегических решений, оптимизации маркетинговых кампаний, разработке новых продуктов и услуг, а также в формировании эффективной коммуникации. В цифровой среде, когда информация генерируется, передаётся и хранится в огромных объемах, анализ предпочтений аудитории через цифровые каналы взаимодействия является основополагающим элементом успешного ведения экономической деятельности и достижения целей в любой сфере.

Разработанный алгоритм состоит из нескольких этапов:

1. Обработка массива содержательных токенов. Основной задачей данного этапа является извлечение значимых для исследования токенов из общего массива информации. Особую ценность будут представлять единицы, связанные с предприятием, его деятельностью, репутацией, взаимосвязям с внешней средой.

2. Формирование и обработка первого поколения кластеров. В рамках второго этапа основной задачей является проведение векторизации информации в сформированном массиве данных, что позволит выделить информационные сектора, характеризуемые определёнными наборами сущностно-содержательных концентраторов.

3. Нахождение оптимального параметрического сочетания. Основная задача – найти оптимальное соотношение параметров. При этом необходимо учитывать не только каждый параметр в отдельности, но и сочетание параметров друг с другом. Таким образом, алгоритмика развития должна быть построена на улучшении качества параметрического состава.

Результаты и обсуждения

Рассмотрим подробнее каждый их этапов.

Этап 1. Обработка массива содержательных токенов.

Для получения массива содержательных токенов в первую очередь необходимо сформировать общий базовый массив, например, аккаунтов потенциальных абитуриентов (в апробации инструментария в массив включены аккаунты школьников тестового региона России) [14]. Далее осуществляется поаккаунтный сбор сообществ, которые являются в предлагаемой алгоритмике маркером интересов владельцев аккаунтов.

При этом, содержательно актуальными являются только сообщества определённого диапазона в распределении количества участников. Слишком большие сообщества с широким кругом интересов не позволяют эффективно использовать акторную структуризацию контента. Аналогично, маленькие сообщества, размерность которых сопоставима с количеством акторов, приводят к низкой продуктивности такой структуризации. Для механического расчёта соответствующего диапазона сообществ проводится их линеаризация по числу участников. Наиболее эффективным преобразованием в данном случае является преобразование Бокса-Кокса, формирующее значимый результат при неизвестном типе распределения. Преобразование для последовательности $y = \{y_1, \dots, y_n\}, y_i > 0, i = 1, \dots, n$ определяется как [3] (1):

$$y_i^{\lambda} = \begin{cases} \frac{y_i^{\lambda} - 1}{\lambda}, \text{если } \lambda \neq 0 \\ \log(y_i), \text{если } \lambda = 0 \end{cases}, \quad (1)$$

где λ выбирается в соответствии с максимизацией логарифма правдоподобия.

После линеаризации осуществляется определение содержательного диапазона (range) (2):

$$Range_{min,max} = Median \pm \frac{\sum_{i=1}^{n} |x_i - \mu|^2}{N}, \quad (2)$$

где Median – медиана массива;

N – размер массива;

μ – среднее арифметическое массива.

В определённом содержательном диапазоне сообществ осуществляется сбор текстов постов и их токенизация [10]. Из массива токенов аналогично массиву сообществ извлекается рабочий содержательный диапазон как часть полного распределения [13].

При этом, в общем случае для построения полного распределения может быть использован закон Ципфа (3):

$$C_i = \frac{f_i * r_i}{n}, \quad (3)$$

где f_i – частота вхождения i-го слова в массив;

r_i – ранг частоты для i-го слова;

n – количество слов в массиве.

Такой подход позволяет графически отразить текст на координатной плоскости, где по оси абсцисс – количество вхождения слова в текст, по оси ординат – ранг слова.

То есть, закон Ципфа направлен на прямое построение распределения текста, на основе которо-

го в алгоритме предлагается выявить значимый диапазон токенов.

Далее необходим такой маркер значимости токенов, который позволяет выделять явные сущностно-содержательные компоненты, на основании чего может быть осуществлена векторизация информации. В этих целях используется статистическая мера tf_idf, отражающая значимость токена в обособленной группе, находящейся внутри массива нескольких обособленных групп [4]. То есть, значимость токена будет тем больше, чем чаще он встречается внутри одной группы параллельно с тем, чем реже он встречается в других группах [15]. Расчёт осуществляется по формуле 4:

$$tf_{idf_i}(w_i, d, D) = tf_i * idf_i \quad , \quad (4)$$

где параметр tf (term frequency) отражает значимость оцениваемого слова в конкретной группе внутри массива групп, а idf (invers document frequency) обратный параметр употребляемости слова в массиве групп (то есть, параметр, отвечающий за снижение значимости конкретного слова внутри группы, если оно часто входит в другие группы), которые рассчитываются по формулам (5, 6):

$$tf_i(w_i, D) = \frac{n_{w_i}}{\sum_d n_d} \quad , \quad (5)$$

$$idf_i(w_i, D) = log \frac{|D|}{|d \in D : w_i \in d|} \quad , \quad (6)$$

где: w – слово в группе;

D – общий массив;

d – группы массива;

n_{wi} – количество употреблений слова w в группе d;

n_d – суммарное количество слов в группе.

Таким образом, получаем базовый массив содержательных токенов, распределённый по уникальности токенов и их значимости для сообществ.

Этап 2. Формирование и обработка первого поколения кластеров

В рамках второго этапа основной задачей является проведение векторизации информации в сформированном массиве данных, что позволит выделить информационные сектора, характеризуемые определёнными наборами сущностно-содержательных концентраторов [11].

Так как в состав массива входят очень часто употребляемые типичные слова и крайне редко употребляемые случайные единицы, необходимо для улучшения качества векторизации текстовой информации провести отсеивание. Целевой интерес представляют слова, располагающиеся в информативно-значимой зоне массива. Верхнюю и нижнюю границы эффективной информативности предлагается ограничить коэффициентами усечения выборки: k_1 и k_2.

Непосредственно после обозначения информационно-значимой зоны в ней для очищенного мешка слов осуществляется формирование групп интересов – векторизация информационных секторов. В этих целях используется инструментарий кластерного анализа – по методу LDA [5].

Кластеризация осуществляется с помощью латентного размещения Дирихле (Latent Dirichlet allocation). Основой данного размещения является интеграл Дирихле (названный именем взявшего его Иоганна Дирихле) [7] (7):

$$\int\limits_{x_i \geq 0, \sum_i x_i = 1} \prod_{i=1}^{n} x_i^{a_i - 1} = B(\alpha) = \frac{\Gamma(\alpha_1) \dots \Gamma(\alpha_n)}{\Gamma(\alpha_1 + \dots + \alpha_n + 1)} \quad , \quad (7)$$

где: Г (α_1) – гамма-функция (Г(n)=(n-1)!).

Тогда сформированное распределение Дирихле от данного интеграла – распределение на векторе размера n-1 (x_1, x_1, …, x_{n-1}) (8):

$$p\left(x_1, \dots, x_{(n-1)} \mid \alpha_1, \dots, \alpha_n\right) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{(\alpha_i - 1)} \quad , \quad (8)$$

где: $x_n = 1 - x_1 - \dots - x_{n-1}$; $x_i > 0$.

Основной особенностью данного распределение является то, что построенный по касательной к нему вектор проходит через неотрицательные значения, суммарно составляющие 1. Что позволяет использовать в дальнейших расчётах распределение вероятности, представленное дискретно.

Таким образом, LDA моделирует распределение тем, содержащихся в data set, основываясь на заданном количестве тем (кластеров). Осуществляется построение терм-документной матрицы с n количеством столбцов $\overrightarrow{y_i}$ где для n токенов, характеризующихся y_{ij} частотой вхождения в сущностно-содержательный силуэт сообщества, вектор будет $y_j = \{y_{1j}, y_{2j}, \ldots, y_{mj}\}$ [6]. То есть, частота вхождения представлена как (9):

$$y_{ij} = tf(t_j, g_i), \quad (9)$$

где: g_i – исследуемое сообщество VK, содержащееся в data set (где $i \in \{1, 2, \ldots, m\}$); $tf(t_j, g_i)$ – частота вхождения токена в текстовый силуэт сообщества.

Тогда сама терм-документная матрица выглядит следующим образом (рис. 1):

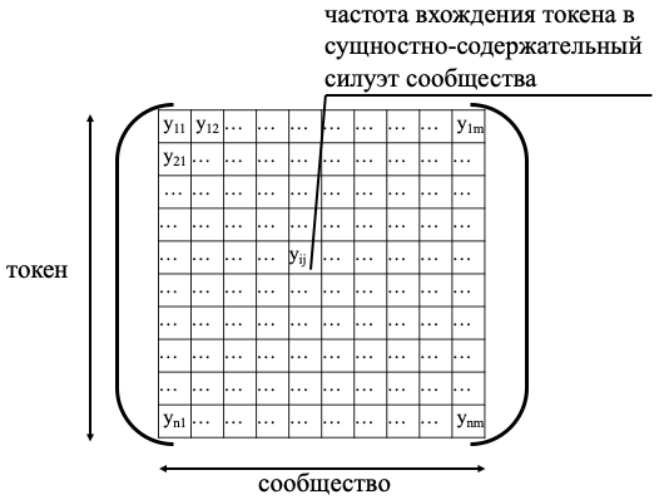

Рис. 1. Терм-документная матрица векторизации сущностно-содержательных компонент data set.
Fig. 1. Term-document matrix of vectorization of the essential semantic components of the data set.

Однако не вся матрица обладает значимостью с точки зрения исследования в контексте сущностно-содержательных компонент. Поэтому, помимо императивно заданного количества кластеров, предлагается в качестве параметров кластеризации также использовать k_1 и k_2, ограничивающие на матричном пространстве информативно-значимую зону. Таким образом, параметрический состав кластеризации на поле терм-документной матрицы представляет собой окно усечения значимых сущностно-содержательных компонент (рис. 2).

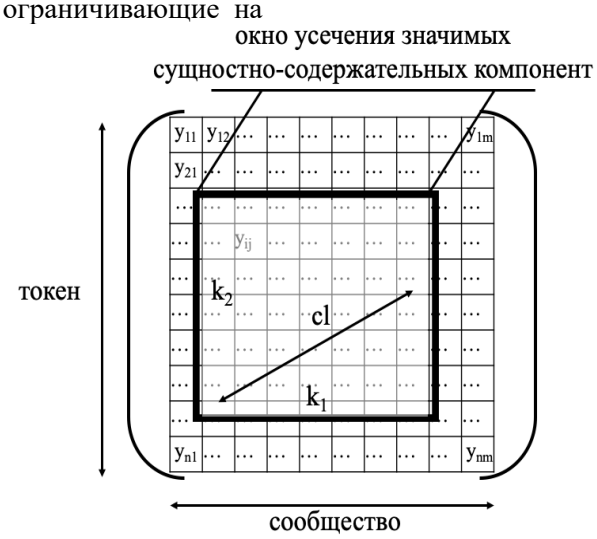

Рис. 2. Наложение окна усечения значимых сущностно-содержательных компонент на базисную терм-документную матрицу.
Fig. 2. Application of the truncation window of significant essential semantic components to the base term-document matrix. Compiled by the authors as part of the conducted research.

Так как осуществляется обработка большого массива, информационные вектора которого заранее не известны, задание количества кластеров и коэффициентов усечения эмпирически не может быть до конца эффективным.

Поэтому значимой аналитической задачей является определение наиболее эффективных численных параметров векторизации информационных секторов.

В первую очередь необходимо установить маркер эффективности секторизации – для этих целей может быть использован коэффициент силуэта. Причём, допускается три различные вариации его внедрения в маркер:

Коэффициент силуэта (классический статистический параметр, используемый без изменения) [8]. При большом количестве кластеров может быть потенциально неэффективен.

Для расчёта классического коэффициента силуэта необходимо определить для каждого i-го элемента каждого кластера среднее расстояние до смежных элементов внутри кластера (Lin_i) и среднее расстояние до ближайших элементов, относящихся к другим кластерам $(Lout_i)$ (10, 11). и среднее расстояние до ближайших элементов, относящихся к другим кластерам

$$Lin_i = \frac{1}{|K_I| - 1} \sum_{f \in K_I, i \neq f} l(i, f) \quad , \qquad (10)$$

где: $|K_I|$ – мощность кластера №1 (количество элементов кластера);

l (i, f) – расстояние от i-го элемента кластера до элемента кластера f.

$$Lout_i = \min_{I \neq A} \frac{1}{|K_A|} \sum_{a \in K_A,} l(i, a) \quad , \qquad (11)$$

где: $\in K_A$ – элемент кластера A, отличного от кластера I.

Тогда коэффициент силуэта для i-го элемента кластера I (12):

$$ks_i = \frac{Lout_i - Lin_i}{\max(Lout_i, Lin_i)} \quad . \qquad (12)$$

Соответственно, чем ближе значение коэффициента силуэта к 1, тем выше качество кластеризации для рассматриваемого элемента кластера. Если значение приближается к -1, для рассматриваемого элемента предпочтительнее соседний кластер, где соседним кластером для элемента считается тот, для которого наименьшее l (i, a).

После определения коэффициента силуэта для каждого элемента кластера, находится среднее значение коэффициента для всех элементов всех кластеров (ks). Данное значение является индикатором качества кластеризации.

Отношение коэффициента силуэта к количеству кластеров (13). Решает проблему метрики. Однако значение количества кластеров в знаменателе дроби формирует линейное влияние, а гипотетически может иметь место быть нелинейность тренда.

$$ks_{weigh} = \frac{ks}{K_{gen}} \quad , \qquad (13)$$

где: K_{gen} – количество кластеров в кластерном размещении.

Отношение коэффициента силуэта к логарифму от количества кластеров (14). Потенциально эффективен, однако будет иметь место зона количества кластеров, в которой аналогично будет выражена прямая зависимость тренда маркера эффективности и тренда увеличения количества кластеров.

$$ks_{wl} = \frac{ks}{logK_{gen}} \quad , \qquad (14)$$

Таким образом, предлагается использовать в качестве маркера эффективности неприведённый коэффициент силуэта. Тогда в целях регулирования его значимости предлагается ввести верхнее ограничение для компоненты, отражающей количество кластеров.

Далее предлагается сформировать первый (базовый) вариант размещения секторов.

Так как эмпирически не может быть осуществлено назначение наиболее эффективных численных параметров векторизации, предлагается для базового варианта размещения определить параметры векторизации случайным образом.

На основании полученных параметров векторизации осуществляется обработка массива данных, в результате чего методологией LDA формируется первое поколение вариантов кластерного размещения, эффективность кластерного состава которых определяется коэффициентом силуэта. Первое поколение содержит 100 вариантов кластерного размещения (vcl).

Этап 3. Нахождение оптимального параметрического сочетания

На данном этапе предлагается сформировать последовательность поколений кластеризации, обладающую трендом увеличения коэффициента силуэта – то есть, создать прогрессивную цепь изменения параметров векторизации. Задача такого плана может быть решена инструментарием, принцип действия которого форсируется на логике

генетики – для каждого поколения выбирается наилучшее сочетание параметров, на основе мутаций от которого строятся параметрические сочетания следующего поколения.

В первую очередь из поколения №1 выбирается 10 лучших (наиболее эффективных по коэффициенту силуэта) вариантов кластерного размещения. Все они в целях сохранения наиболее эффективных параметрических сочетаний без изменений параметрического состава переходят в следующее поколения. А 5 из них становятся базой для мутаций нового поколения.

Интерес, соответственно, представляет параметрическое сочетание данных вариантов. Лучших параметрические сочетания поколения №1 становятся базой мутаций, на основе которых формируются параметрические сочетания поколения №2.

Мутация осуществляется случайным образом в ограниченном коридоре мутации. Каждый параметр с равной долей вероятности может либо сохранить своё значение, либо мутировать на 10% в большую или меньшую сторону. Посредством таких мутаций формируется 90 вариантов кластерного размещения, ещё 10 вариантов нового поколения составляют 10 наилучших вариантов – предшествующего. Это позволяет держать под контролем процессы случайных мутаций и формирует положительный тренд развития коэффициента силуэта из поколения в поколение.

Алгоритмика генетического формирования второго поколения кластерных размещений представлена на рис. 3.

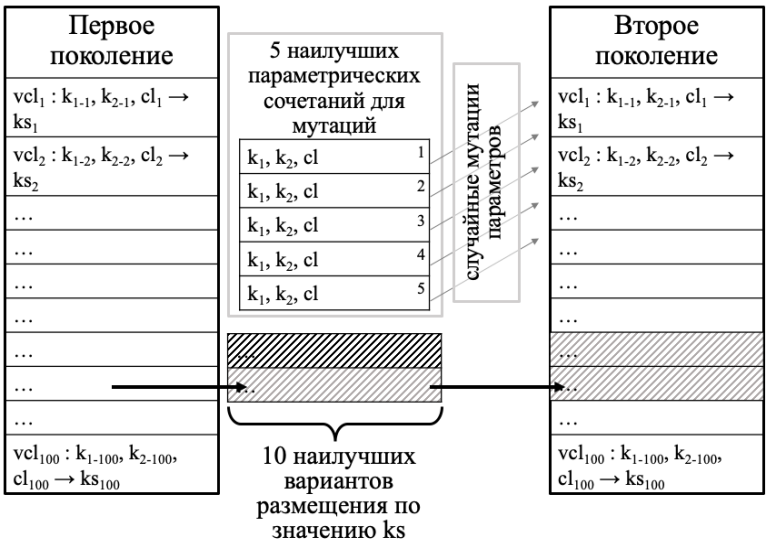

Рис. 3. Генетический алгоритм формирования параметрических сочетаний кластеризации второго и последующих поколений.

Fig. 3. Genetic algorithm for generating parametric combinations of clustering for the second and subsequent generations.

Далее проводятся итерации алгоритмики в последовательности формирования поколений. Она осуществляется до тех пор, пока коэффициент силуэта не выйдет на плато эффективности (или не достигнет достаточного уровня эффективности в рамках целеполагания). Схематично логика параметрического состава для плато эффективности представлена на рис. 4.

Рис. 4. Параметрическое содержание плато эффективности. Составлено авторами в рамках проведенного исследования.
Fig. 4. Parametric composition of the efficiency plateau. Compiled by the authors as part of the conducted research.

Поэтапно общий алгоритм инструментария представлен на рис. 5-6.

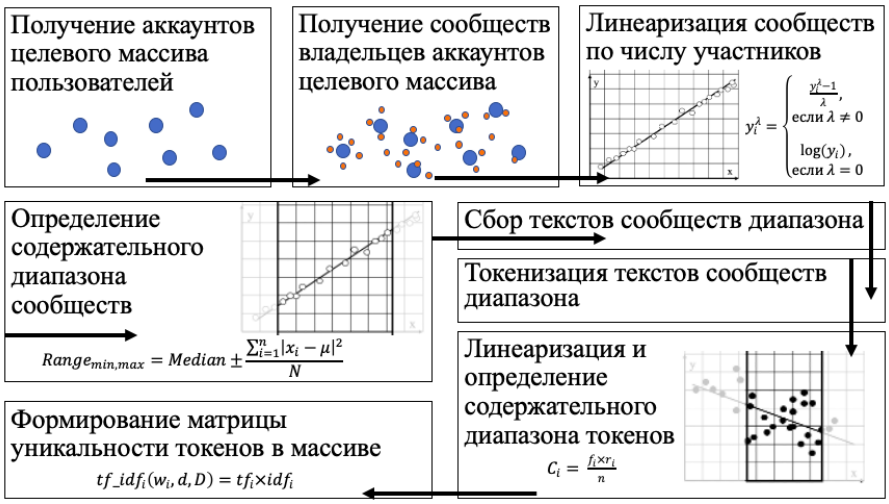

Рис. 5. 1 этап работы алгоритмики: формирование массива содержательных токенов.
Fig. 5. Stage 1 of the algorithm operation: formation of the array of meaningful tokens.

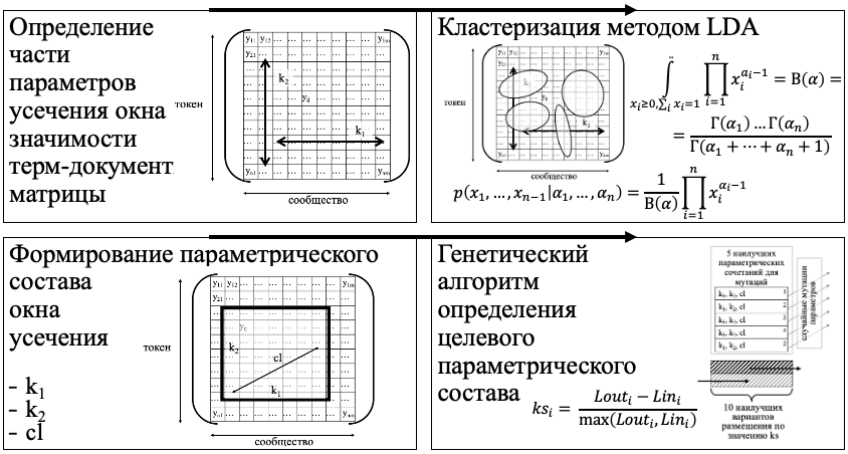

Рис. 6. 2 и 3 этапы работы алгоритмики: формирование состава параметров и получение оптимального параметрического сочетания.
Fig. 6. Stages 2 and 3 of the algorithm operation: formation of the parameter set and obtaining the optimal parametric combination.

Таким образом, формируется эффективное сочетание параметров. Ключевой наилучший полученный результат представляет собой распределение сущностно-содержательных компонент по кластерам, формирующее характеристики групп интересов. Данный результат, агрегируемая им информация, может быть использован для решения широкого круга задач.

Он позволяет сегментировать целевую аудиторию на группы с похожими интересами и предпочтениями. Это даёт возможность оптимизировать коммуникационные стратегии, направляя информацию и предложения на каждый кластер в соответствии с его особенностями. Например, кластер людей, интересующихся здоровьем, будет реагировать на рекламу диет и фитнес-программ, в то время как кластер любителей игр будет более восприимчив к рекламе новинок в игровой индустрии.

Распределение сущностно-содержательных компонент по кластерам позволяет понять структуру и динамику целевой аудитории. Анализ характеристик каждого кластера позволяет определить ключевые факторы, влияющие на интересы и поведение группы. Это позволяет прогнозировать будущие изменения в потребительском поведении и разрабатывать стратегии адаптации к ним.

Дифференцированные характеристики групп интересов могут быть использованы для разработки новых продуктов и услуг. Изучение интересов и потребностей разных кластеров позволяет определить незаполненные ниши на рынке и разработать предложения, специально направленные на эти ниши.

Таким образом, инструмент позволяет определить структуру интересов целевой аудитории и преобразовать разрозненные данные о предпочтениях в значения, агрегированные в формате, удобном для анализа. Это открывает новые возможности для эффективной таргетированной коммуникации, разработки инновационных продуктов и услуг, а также для понимания и прогнозирования изменений в потребительском поведении.

Алгоритм, предложенный в данном исследовании имеет широкую сферу применения и представляет ценность для владельцев предприятия. В первую очередь, это упрощает анализ целевой аудитории продукта и формирование представления об ее отношении к бренду. Следовательно, выделение интересов клиента позволяет предлагать персонализированные предложения и контент, повысить лояльность и стимулировать повторные покупки. Во-вторых, это позволит выделить тренды в потребительском поведении и своевременно внести изменения в продукте. Особенно актуален данный аспект при разработке инновационных продуктов и технологичных решений, где высокий уровень конкуренции. В третьих, анализ информационной среды предприятия дает шанс рассмотреть продукт со стороны клиента, то есть обнаружить определенные характеристики продукта, за которые он готов платить. Такая информация позволит пересмотреть политику ценообразования и установить оптимальную цену, максимизируя прибыль и сохраняя конкурентоспособность.

Выводы

В исследовании построен инструментарий определения групп интересов для выбранного массива аккаунтов пользователей в цифровой ин-

формационной среде, сформированного по целевым параметрам. Инструментарий автоматизирует математическую и аналитическую обработку текстовой информации, генерируемой целевой группой людей.

Обобщаются стандартизированные ступени выявления сущностно-содержательных компонент в текстовом массиве, формирующие выборку основных информационных составляющих текста, уникальность которых является качественным параметром кластеризации.

Данная алгоритмика позволяет дифференцировать людей по кластерам, отражающим их интересы, относящиеся к качественно значимым с точки зрения таргетированного взаимодействия с владельцами аккаунтов: диапазон популярности интересов выбран таким образом, чтобы не учитывать наиболее общие и наиболее локальные интересы, которые не могут быть эффективны в дифференциации уникальных пользовательских предпочтений.

То есть, методология значима как завершённый результат с точки зрения анализа. Кроме того, в качестве промежуточного этапа разработанный инструмент может использоваться для построения алгоритмов формирования/поддержания/развития целевого тематического интереса групп людей, сущностным базисом которого будет являться сочетание паттернов интереса с паттернами целевой тематики. Его можно использовать для таргетирования маркетинговых кампаний, при определении групп людей с высокой вероятностью интереса к конкретной продукции или услуге. Сочетание паттернов интереса с паттернами целевой тематики позволяет сегментировать аудиторию и создавать персонализированные рекламные сообщения, повышая эффективность кампаний. Также инструмент может применяться в образовательной сфере, где он может помочь в разработке персонализированных учебных программ и подборе материалов, соответствующих индивидуальным интересам студентов. Анализ паттернов интереса позволяет выявить склонности и предпочтения, что дает возможность сделать обучение более эффективным и мотивированным. Внедрение модуля тематического углубления по направлениям будущей специализации в образовательные программы довузовской подготовки высших учебных заведений представляет собой перспективную стратегию повышения мотивации абитуриентов и увеличения их притока. Также стоит отметить, что на основе данных об интересах абитуриентов можно проводить более эффективные профориентационные консультации, помогая им сделать осознанный выбор будущей специальности. Кроме того, инструмент может быть полезен в социальных исследованиях для анализа общественного мнения и выявления ключевых тематических интересов различных групп населения, позволяя изучать структуру и динамику общественного мнения, выявлять тренды и прогнозировать возможные изменения. Аналогично в политическом маркетинге – инструмент позволяет фиксировать группы людей с определенными политическими убеждениями и интересами, способствуя оптимизации политических кампании.

В целом, представленный инструмент обладает потенциалом для применения в различных областях, связанных с формированием целевых групп и персонализацией контента. Например, для предприятия важно своевременно адаптировать продукт, особенно в высоких конкурентных технологичных сферах, минимизируя риски убытков, выявлять характеристики продукта, ценные для клиента, и устанавливать оптимальную цену для максимизации прибыли и конкурентоспособности.

Существенным преимуществом разработанного алгаритма является его масштабируемость и эффективность: он позволяет эффективно обрабатывать большие объемы данных о пользователях и масштабировать анализ на пользователей. Также стоит отметить, что акторное доминирование генов позволяет повысить релевантность полученных результатов и учитывать значимаю для исследования информацию. Для бизнеса это позволит найти связи, которые не очевидны при ручном анализе.

Финансирование

Работа выполнена в рамках реализации проекта «Разработка методологии формирования инструментальной базы анализа и моделирования пространственного социально-экономического развития систем в условиях цифровизации с опорой на внутренние резервы» (FSEG-2023-0008)

Список источников

1. Rainie H., Wellman B. Networked: the new social operating system. Cambridge, MA: MIT Press, 2012. T. 10.

2. Horsmann T., Zesch T. LTL-UDE EmpiriST 2015: tokenization and PoS tagging of social media text. Proceedings of the 10th Web as Corpus Workshop. 2016. P. 120 – 126.

3. Box G.E.P., Cox D.R. An analysis of transformations // Journal of the Royal Statistical Society Series B: Statistical Methodology. 1964. T. 26. No. 2. P. 211 – 243.

4. Singh S., Kumar K., Kumar B. Sentiment analysis of Twitter data using TF-IDF and machine learning techniques. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) // IEEE, 2022. Vol. 1. P. 252 – 255.

5. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. T. 3. P. 993 – 1022.

6. Бондарчук Д.В. Алгоритмы интеллектуального поиска на основе метода категориальных векторов: специальность 05.13.17 «Теоретические основы информатики»: дис. … канд. физ.-матем. наук. Екатеринбург: Уральский государственный университет путей сообщения, 2016. 141 с.

7. Richter S., Sundberg C. A formula for the local Dirichlet integral // Michigan Mathematical Journal. 1991. Vol. 38. No. 3. P. 355 – 379.

8. Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53 – 65.

9. Rodionov D.G., et al. Information environment quantifiers as investment analysis basis // Economies. 2022. Vol. 10. No. 10. P. 232 – 254.

10. Rodionov D.G., et al. The information environment cluster distribution of the regional socio-economic systems in transition economy. International Scientific Conference “Digital Transformation on Manufacturing, Infrastructure & Service”. Cham: Springer Nature Switzerland, 2022. P. 203 – 217.

11. Rodionov D.G., et al. Modeling changes in the enterprise information capital in the digital economy // Journal of Open Innovation: Technology, Market, and Complexity. 2021. Vol. 7. No. 3. P. 166 – 189.

12. Rodionov D.G., et al. A new model for creating digital avatars of applicants based on social media data // International Journal of eBusiness and eGovernment Studies. 2023. T. 15. No. 1. P. 324 – 341.

13. Rodionov D.G., et al. Modeling a digital avatar of car drivers based on the quantification of the information environment // International Conference on Next Generation Wired/Wireless Networking. Cham: Springer Nature Switzerland. 2023. P. 170 – 180.

14. Родионов Д.Г., Конников Е.А., Конникова О.А., и др. Влияние социально-экономических факторов на проникновение интернет-маркетинга в различных странах // Мягкие измерения и вычисления. 2024. № 1. С. 48 – 61.

15. Родионов Д.Г., Конников Е.А., Пашинина П.А., Шаныгин С.И. Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF. Мягкие измерения и вычисления. 2024. № 3. С. 72 – 84.

References

1. Rainie H., Wellman B. Networked: the new social operating system. Cambridge, MA: MIT Press, 2012. Vol. 10.

2. Horsmann T., Zesch T. LTL-UDE EmpiriST 2015: tokenization and PoS tagging of social media text. Proceedings of the 10th Web as Corpus Workshop. 2016. P. 120 – 126.

3. Box G.E.P., Cox D.R. An analysis of transformations. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1964. T. 26. No. 2. P. 211 – 243.

4. Singh S., Kumar K., Kumar B. Sentiment analysis of Twitter data using TF-IDF and machine learning techniques. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON). IEEE, 2022. Vol. 1. P. 252 – 255.

5. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003. Vol. 3. P. 993 – 1022.

6. Bondarchuk D.V. Intelligent search algorithms based on the categorical vector method: specialty 05.13.17 "Theoretical Foundations of Computer Science": dissertation for the degree of candidate of physical and mathematical sciences. Ekaterinburg: Ural State University of Railway Engineering, 2016. 141 p.

7. Richter S., Sundberg C. A formula for the local Dirichlet integral. Michigan Mathematical Journal. 1991. T. 38. No. 3. P. 355 – 379.

8. Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53 – 65.

9. Rodionov D.G., et al. Information environment quantifiers as investment analysis basis. Economies. 2022. Vol. 10.No. 10. P. 232 – 254.

10. Rodionov D.G., et al. The information environment cluster distribution of the regional socio-economic systems in transition economy. International Scientific Conference “Digital Transformation on Manufacturing, Infrastructure & Service”. Cham: Springer Nature Switzerland, 2022. P. 203 – 217.

11. Rodionov D.G., et al. Modeling changes in the enterprise information capital in the digital economy. Journal of Open Innovation: Technology, Market, and Complexity. 2021. Vol. 7.No. 3. P. 166 – 189.

12. Rodionov D.G., et al. A new model for creating digital avatars of applicants based on social media data. International Journal of eBusiness and eGovernment Studies. 2023. T. 15. No. 1. P. 324 – 341.

13. Rodionov D.G., et al. Modeling a digital avatar of car drivers based on the quantification of the information environment. International Conference on Next Generation Wired/Wireless Networking. Cham: Springer Nature Switzerland. 2023. P. 170 – 180.

14. Rodionov D.G., Konnikov E.A., Konnikova O.A., et al. The influence of socio-economic factors on the penetration of Internet marketing in different countries. Soft measurements and calculations. 2024. No. 1. P. 48 – 61.

15. Rodionov D.G., Konnikov E.A., Pashinina P.A., Shanygin S.I. Thematic modeling of the information environment of media companies: the LDA-TF-IDF toolkit. Soft measurements and calculations. 2024. No. 3. P. 72 – 84.

Информация об авторах

Панкова Л.В., кандидат экономических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Новороссийская ул., 50, pankova_lv@spbstu.ru

Старченкова О.Д., ассистент, специалист учебной лаборатории «Моделирование и цифровизация социально-экономических систем», Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Новороссийская ул., 50